Extracting Structural Information from Residual Chemical Shift Anisotropy: Analytic Solutions for Peptide Plane Orientations and Applications to Determine Protein Structure*

Chittaranjan Tripathy[1], Anthony K. Yan[1,2], Pei Zhou[2], and Bruce Randall Donald[1,2,**]

[1] Department of Computer Science, Duke University, Durham, NC 27708
[2] Department of Biochemistry, Duke University Medical Center, Durham, NC 27710
brd+recomb13@cs.duke.edu

Abstract. Residual dipolar coupling (RDC) and residual chemical shift anisotropy (RCSA) provide orientational restraints on internuclear vectors and the principal axes of chemical shift anisotropy (CSA) tensors, respectively. Mathematically, while an RDC represents a single spheroconic, an RCSA can be interpreted as a *linear combination* of *two* spheroconics. Since RDCs and RCSAs are described by a molecular alignment tensor, they contain inherent structural ambiguity due to the symmetry of the alignment tensor and the symmetry of the molecular fragment, which often leads to more than one orientation and conformation for the fragment consistent with the measured RDCs and RCSAs. While the orientational multiplicities have been long studied for RDCs, structural ambiguities arising from RCSAs have not been investigated. In this paper, we give exact and tight bounds on the number of peptide plane orientations consistent with multiple RDCs and/or RCSAs measured in one alignment medium. We prove that at most 16 orientations are possible for a peptide plane, which can be computed in closed form by solving a merely quadratic equation, and applying symmetry operations. Furthermore, we show that RCSAs can be used in the initial stages of structure determination to obtain highly accurate protein backbone global folds. We exploit the mathematical interplay between spheroconics derived from RCSA and RDC, and protein kinematics, to derive quartic equations, which can be solved in closed-form to compute the protein backbone dihedral angles (ϕ, ψ). Building upon this, we designed a novel, sparse-data, polynomial-time divide-and-conquer algorithm to compute protein backbone conformations. Results on experimental NMR data for the protein human ubiquitin demonstrate that our algorithm computes backbone conformations with high accuracy from $^{13}C'$-RCSA or ^{15}N-RCSA, and N-H^N RDC data. We show that the structural information present in $^{13}C'$-RCSA and ^{15}N-RCSA can be extracted analytically, and used in a rigorous algorithmic framework to compute a high-quality protein backbone global fold, from a limited amount of NMR data. This will benefit automated NOE assignment and high-resolution protein backbone structure determination from sparse NMR data.

* Grant sponsor: National Institutes of Health; Grant numbers: R01 GM-65982 and R01 GM-78031 to BRD, and R01 GM-079376 to PZ

** Corresponding author

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is one of the most powerful experimental techniques for the study of macromolecular structure and dynamics, particularly for proteins in solution. NMR complements X-ray crystallography in that it can obtain structural information for proteins that are hard to crystallize, intrinsically disordered proteins [18,38], and denatured proteins [36]. NMR has also emerged as a major tool to probe protein-ligand interactions [20] under near physiological conditions, as well as to investigate invisible excited states in proteins and extract information on these minor conformers [3, 4].

The NMR technique is based on the sensitivity of magnetic properties of the nuclei to its local chemical and electronic environment in the presence of a strong and static external magnetic field of the spectrometer. The observable for a nucleus, called its *chemical shift*, arises from the nuclear shielding effect caused by the local magnetic field, induced by the circulation of electrons surrounding the nucleus. This induced field can be described by a second-rank chemical shift (or shielding) anisotropy (CSA) tensor, which can be rewritten to correspond to the isotropic, anisotropic antisymmetric, and anisotropic symmetric parts.

In solution, due to isotropic molecular tumbling, the anisotropic parts of the CSA tensor average out to zero due to rotational diffusion, and only the remaining isotropic chemical shift, $\delta_{\mathtt{iso}}$, is observed. While isotropic chemical shifts play an increasingly important role in NMR structure elucidation and refinement [9, 45,59], and dynamics [1], our understanding of the relationship between structure and chemical shifts is still far from complete, especially in the context of proteins [68] and other macromolecules. The antisymmetric part of the CSA tensor often has a negligible effect on the relaxation rates, and hence can be ignored. The symmetric part of the CSA tensor is a traceless, second-rank tensor, usually represented by its three eigenvalues in the *principal order frame* and the orientations of its three principal axes (eigenvectors) with respect to the molecular frame.

Accurate knowledge of CSA tensors is essential to the quantitative determination and interpretation of dynamics, relaxation interference [33, 46], residual chemical shift anisotropy (RCSA) [12,39,63], and NMR structure determination and refinement [15, 26, 27, 61]. In solid-state NMR, CSA tensors can be determined from powder patterns [32], or magic-angle spinning (MAS) spectra [60]. In solution NMR, the CSA tensors can be determined from relaxation and CSA-dipolar cross-correlation experiments [30, 62, 63], or from offsets in resonance peaks upon partial alignment [7]. The presence of an alignment medium introduces partial alignment in the molecules. Residual dipolar coupling (RDC) [34] can easily be extracted, often with high-precision, as the difference between the line-splittings in weakly aligned and isotropic buffer solutions. The small difference in chemical shifts observed under partially aligned conditions and isotropic conditions gives rise to the RCSA effect [39]. Techniques to measure RCSA include temperature-dependent phase transition of certain liquid crystals [12], varying concentration of the aligning medium [6], utilizing MAS to eliminate the effects of protein alignment relative to the magnetic field [21], and the recently introduced two-stage NMR tube method by Prestegard and coworkers [28].

Similar to RDC, RCSA contains rich and orientationally sensitive structural information [46] that complements other types of structural restraints such as the nuclear Overhauser effect (NOE) distance restraints and scalar couplings. Since amide nitrogen RCSA (^{15}N-RCSA) and carbonyl RCSA (^{13}C′-RCSA) can be measured to high precision [12, 63], they have been used as structural restraints for protein structure validation [14] and refinement [10,26,27,53] during the *final* stages of traditional protein structure determination [5, 42]. However, to our knowledge, RCSAs have never been used in the *initial* stages of structure computation to compute the backbone global fold of a protein. Methods that primarily use RDCs in the initial stages of structure computation [16, 19, 47, 65] have been shown to have many advantages over traditional NOE-based structure determination protocols. Recently, in [35, 45], Baker and Bax and coworkers have developed protocols within the ROSETTA [25] protein structure modeling framework, that use only backbone chemical shifts, RDCs, and amide proton NOE distances to compute high-quality protein backbone conformations. However, these approaches do not use structural restraints from RCSA data. Further, most of these approaches use stochastic search, and therefore, lack any algorithmic guarantee on the quality of the solution or running time.

In recent work from our laboratory [17, 51, 52, 56, 58, 64], polynomial-time algorithms have been proposed for high-resolution backbone global fold determination from a minimal amount of RDC data. This framework is called RDC-ANALYTIC. The core of the RDC-ANALYTIC suite is based on representing RDC and protein kinematics in algebraic form, and solving them analytically to obtain closed-form solutions for the backbone dihedrals and peptide plane orientations, in a divide-and-conquer framework to compute the global fold. These algorithms have been used in [57, 65, 66] to develop new algorithms for NOE assignment, which led to the development of a new framework [65] for high-resolution protein structure determination, which was used prospectively to solve the solution structure of the FF Domain 2 of human transcription elongation factor CA150 (FF2) (PDB id: 2KIQ). Recently, we have developed a novel algorithm, POOL [51, 52], within the RDC-ANALYTIC framework, to determine protein loop conformations from a minimal amount of RDC data. However, RDC-ANALYTIC did not exploit orientational restraints from RCSA data.

In this work, we show that orientational restraints from ^{13}C′-RCSAs or ^{15}N-RCSAs can be used in combination with N-H$^{\mathrm{N}}$ RDCs in an analytic, systematic search-based, divide-and-conquer framework to determine individual peptide plane orientations and protein backbone conformations. Our new algorithm is a part of the RDC-ANALYTIC framework, and is called RDC-CSA-ANALYTIC. Two demonstrations of applying RDC-CSA-ANALYTIC, (1) using ^{13}C′-RCSA and N-H$^{\mathrm{N}}$ RDC, and (2) using ^{15}N-RCSA and N-H$^{\mathrm{N}}$ RDC, to compute the global fold of ubiquitin, and promising results from the application of our algorithm on real biological NMR data, are presented below.

Furthermore, we pursued the fundamental question of determining the peptide plane orientations when 3 measurements are used, each of which is either an RCSA on a nucleus or an RDC on an internuclear vector on the peptide

plane. This is important, because for perdeuterated proteins, RDCs are usually measured on N-H^N, C^α-C', and C'-N coplanar vectors. Further, ^{15}N-RCSA and $^{13}C'$-RCSA can be interpreted with respect to the CSA tensor components on the peptide plane. Previously, Brüschweiler and coworkers [23] showed that it is possible to derive analytic expressions, containing transcendental functions, for the 16 possible peptide plane orientations using only RDCs. However, they only showed a lower bound on the number of solutions. In addition, their work did not consider orientational restraints from RCSAs. In this work, we derive closed-form analytic expressions for the peptide plane orientations from RCSAs and RDCs on coplanar vectors measured in one alignment medium. We prove that *at most* 16 orientations are possible for the peptide plane, which can be computed in closed form by solving a *quadratic equation*, and then applying symmetry operations. This is remarkable because for decades, all previous approaches required, at worst, solving equations involving transcendental functions, or at best, solving polynomial equations of degree 4 or higher. We give a $\Theta(1)$-time deterministic algorithm, 3PLANAR, to compute all possible peptide plane orientations.

2 Theory and Methods

2.1 Residual Dipolar Coupling

The residual dipolar coupling r between two spin-$\frac{1}{2}$ nuclei a and b, described by a unit internuclear vector $\mathbf{v}$, due to anisotropic distribution of orientations in the presence of an alignment medium, relative to a strong static magnetic field direction $\mathbf{B}$ is given by [16, 48, 49]

$$r = D_{\max}\mathbf{v}^T\mathbf{S}\mathbf{v}. \tag{1}$$

Here the dipolar interaction constant $D_{\max}$ depends on the gyromagnetic ratios of the nuclei a and b, and the vibrational ensemble-averaged inverse cube of the distance between them. $\mathbf{S}$ is the *Saupe order matrix* [40], or *alignment tensor* that specifies the ensemble-averaged anisotropic orientation of the protein in the laboratory frame. $\mathbf{S}$ is a 3×3 symmetric, traceless, rank 2 tensor with five independent elements [34, 48, 49]. Letting $D_{\max} = 1$ (i.e., scaling the RDCs appropriately), and considering a global coordinate frame that diagonalizes $\mathbf{S}$, often called the *principal order frame* (POF), Eq. (1) can be written as

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2, \tag{2}$$

where S_{xx}, S_{yy} and S_{zz} are the three diagonal elements of a diagonalized alignment tensor $\mathbf{S}$, and x, y and z are, respectively, the x, y and z components of the unit vector $\mathbf{v}$ in a POF that diagonalizes $\mathbf{S}$. Note that $S_{xx} + S_{yy} + S_{zz} = 0$ because $\mathbf{S}$ is traceless. Since $\mathbf{v}$ is a unit vector, an RDC constrains the corresponding internuclear vector $\mathbf{v}$ to lie on the intersection of a concentric unit sphere and a quadric (Eq. (2)). This gives a pair of closed curves inscribed on the unit sphere that are diametrically opposite to each other (see Figure 1 (a, b)). These curves are known as *sphero-conics* or *sphero-quartics* [8,37]. Since $|\mathbf{v}| = 1$, Eq. (2) can be rewritten in the following form:

$$ax^2 + by^2 = c, \tag{3}$$

where $a = S_{xx} - S_{zz}$, $b = S_{yy} - S_{zz}$, and $c = r - S_{zz}$. Henceforth, we refer to Eq. (3) as the *reduced RDC equation*. For further background on RDCs and RDC-based structure determination, the reader is referred to [16, 17, 34, 48, 49].

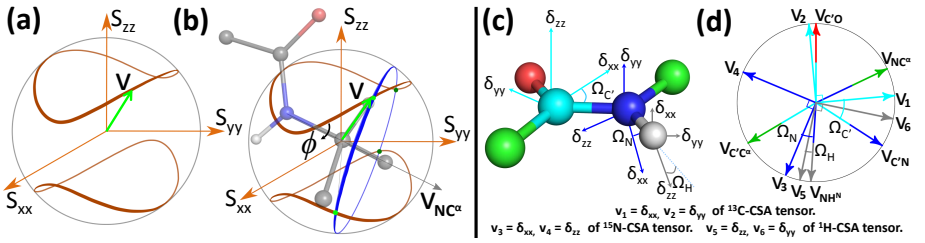

Fig. 1: *Left Panel.* (a) The internuclear vector $\mathbf{v}$ (green arrow) is constrained to lie on one of the two pringle-shaped RDC sphero-conic curves (brown) lying on a unit sphere. (b) The kinematic circle (blue), of the internuclear vector $\mathbf{v}$ (here $\mathbf{v}_{\mathrm{C}^\alpha\mathrm{H}^\alpha}$), around the axis $\mathbf{v}_{\mathrm{NC}^\alpha}$, intersects the sphero-conic curves in at most four points (green dots) leading to a maximum of four possible orientations for $\mathbf{v}$. The case is similar when ϕ is solved and a ψ-defining RDC is measured for an internuclear vector $\mathbf{v}$, e.g., $\mathbf{v}_{\mathrm{NH^N}}$. *Right Panel.* (c) Orientations of the principal components of ^{13}C-, ^{15}N- and ^{1}H-CSA tensors with respect to the peptide plane are shown in cyan, blue and gray, respectively. δ_{zz} is the most- and δ_{xx} is the least-shielded component. For each tensor, one of the components is approximately perpendicular to the peptide plane; therefore, the other two components lie on the peptide plane, and are completely defined by the angle $\mathbf{\Omega}$. The values of the angles $\mathbf{\Omega}_{\mathrm{C}'}$, $\mathbf{\Omega}_{\mathrm{N}}$ and $\mathbf{\Omega}_{\mathrm{H}}$ can be set to fixed values [10, 12], e.g., 38°, 19° and 8° as reported in [12]. (d) The wagon wheel shows the CSA tensor components on the peptide plane along with the bond vectors drawn using C′ atom as the origin.

2.2 Residual Chemical Shift Anisotropy

For a given nucleus, the difference in chemical shifts between the liquid crystalline phase ($\delta_{\mathtt{aniso}}$) and the isotropic phase ($\delta_{\mathtt{iso}}$) is the RCSA, and is given by [12, 14, 39, 63]

$$\Delta\delta = \delta_{\mathtt{aniso}} - \delta_{\mathtt{iso}} = \frac{2}{3} \sum_{i\in\{x,y,z\}} \langle P_2(\cos\theta_{ii})\rangle\, \delta_{ii}, \tag{4}$$

where $P_2(\alpha) = (3\alpha^2 - 1)/2$ is the second Legendre polynomial, δ_{xx}, δ_{yy} and δ_{zz} are the principal components of the CSA tensor, and θ_{xx}, θ_{yy} and θ_{zz} are the respective angles between the principal axes of the traceless, second-rank CSA tensor and the magnetic field direction $\mathbf{B}$. The angle brackets, $\langle\cdots\rangle$, denote ensemble averaging. After suitable algebraic manipulations, we can write Eq. (4) as

$$\Delta\delta = \lambda_1 \boldsymbol{\delta}_{xx}^T \mathbf{S} \boldsymbol{\delta}_{xx} + \lambda_2 \boldsymbol{\delta}_{yy}^T \mathbf{S} \boldsymbol{\delta}_{yy}, \tag{5}$$

where

$$\lambda_1 = \frac{1}{3}(2\delta_{xx} + \delta_{yy}) \tag{6}$$

and

$$\lambda_2 = \frac{1}{3}(2\delta_{yy} + \delta_{xx}) \tag{7}$$

are two constants since δ_{xx} and δ_{yy} are known experimentally. Eq. (5) therefore expresses the $\Delta\delta$ as a *linear combination* of two virtual RDC sphero-conics on two unit vectors $\boldsymbol{\delta}_{xx}$ and $\boldsymbol{\delta}_{yy}$ that can be realized on the peptide plane. This derivation applies, *mutatis mutandis,* to any two choices of unit vectors from $\{\boldsymbol{\delta}_{xx}, \boldsymbol{\delta}_{yy}, \boldsymbol{\delta}_{zz}\}$. Working in the POF of the molecular alignment tensor, we can write the above equation as

$$\Delta\delta = S_{xx}(\lambda_1 x_1^2 + \lambda_2 x_2^2) + S_{yy}(\lambda_1 y_1^2 + \lambda_2 y_2^2) + S_{zz}(\lambda_1 z_1^2 + \lambda_2 z_2^2), \tag{8}$$

where the unit vectors $\boldsymbol{\delta}_{xx} = (x_1, y_1, z_1)^T$ and $\boldsymbol{\delta}_{yy} = (x_2, y_2, z_2)^T$ in the POF of the molecular alignment tensor. Eq. (8) can be simplified to the following form

Table 1: RDC-CSA-ANALYTIC uses a ϕ-defining RDC to compute the backbone dihedral ϕ, and a ψ-defining RDC or RCSA to compute the backbone dihedral ψ exactly and in closed form.

ϕ-defining RDC	C^{α}-H^{α}, C^{α}-C', C^{α}-C^{β}
ψ-defining RDC/RCSA	N-H^{N}, C'-N, C'-H^{N}, $^{13}C'$-RCSA, ^{15}N-RCSA, ^{1}H-RCSA

$$a(\lambda_1 x_1^2 + \lambda_2 x_2^2) + b(\lambda_1 y_1^2 + \lambda_2 y_2^2) = c, \tag{9}$$

where $a = S_{xx} - S_{zz}, b = S_{yy} - S_{zz}$, and $c = \Delta\delta - (\lambda_1 + \lambda_2)S_{zz}$. Henceforth, we refer to Eq. (9) as the *reduced RCSA equation*.

Figure 1 (c, d) shows the local structure of a peptide plane on which the principal components of ^{13}C-, ^{15}N- and ^{1}H-CSA tensors are realized. δ_{zz} and δ_{xx} are respectively the most- and least- shielded CSA tensor components. We denote $^{13}C'$-RCSA, ^{15}N-RCSA and ^{1}H-RCSA by $\Delta\delta_{C'}$, $\Delta\delta_{N}$ and $\Delta\delta_{H}$, respectively.

2.3 The RDC-CSA-ANALYTIC Algorithm

RDC-CSA-ANALYTIC computes the backbone global fold of proteins using RDC and RCSA data in one alignment medium. Table 1 describes the RDC and RCSA types that RDC-CSA-ANALYTIC uses to compute the backbone dihedrals exactly and in closed form. A *ϕ-defining* RDC is used to compute the backbone dihedral ϕ, and a *ψ-defining* RDC or RCSA is used to compute the backbone dihedral ψ, in the increasing order of residue number. The input data to RDC-CSA-ANALYTIC include: (1) the primary sequence of the protein; (2) any combination of at least two RDCs or RCSAs per residue measured in one alignment medium; (3) a sparse set of NOEs; (4) secondary structure element (SSE) boundaries based on TALOS [13, 44] dihedral restraints; and (5) the rotamer library [31].

Previously, we have shown that when a ϕ-defining and a ψ-defining RDC are available for a residue, the corresponding values for ϕ and ψ can be computed by solving quartic equations [51, 52, 64]. RDC-CSA-ANALYTIC extends this to the cases when a ψ-defining RCSA is available in addition to a ϕ-defining RDC (see Proposition 1 below), e.g., when C^{α}-C' or C^{α}-H^{α} RDC, and $^{13}C'$-RCSA or ^{15}N-RCSA data is available. However, in solution NMR, $^{13}C'$-RCSA and/or ^{15}N-RCSA can be measured, often with high precision, along with N-H^{N} RDC, for large and perdeuterated protein systems, for which C^{α}-H^{α} RDCs at the chiral C^{α} center cannot be measured, and C^{α}-C' RDCs measurements are often less precise. Therefore, it is important to be able to determine the global fold from these types of measurements. RDC-CSA-ANALYTIC algorithm specifically provides a solution to this problem. Here we solve the most general case when two ψ-defining RCSAs and/or RDCs are available for residues. Further, this includes the case when (only) two RCSAs per residue are available. It can be shown that (see the supporting information (SI) **Appendix A** available online [50]) one must solve a 32 degree univariate polynomial equation to solve for all possible (at most 32) (ϕ, ψ) pairs, which is a difficult computational problem.

However, for a given value of ϕ_i, the values of ψ_i can be computed by solving a quartic equation (see Proposition 1 below). Here we present a hybrid approach that employs a systematic search over ϕ combined with solutions to two quartic

equations for ψ derived from two ψ-defining RDC/RCSA values r_1 and r_2, to compute the backbone dihedrals (ϕ, ψ) pairs. For each ϕ, sampled systematically from the Ramachandran map, let A and B (each of size $\leq$4) be the sets of all ψ values computed using r_1 and r_2, respectively. If $A \cap B \neq \emptyset$, then for a $\psi \in A \cap B$, the corresponding (ϕ, ψ) pair is a solution. However, in practice, there are two issues that need to be addressed. First, due to finite-resolution sampling of ϕ, and experimental errors in the RDC and RCSA data, the intersection of sets A and B can be an empty set, even though there exist $\psi_A \in A$ and $\psi_B \in B$ such that $|\psi_A - \psi_B| < \delta$, for some small delta $\delta > 0$ which depends on the resolution of sampling of ϕ. This issue can be addressed by choosing a suitable resolution α for systematic sampling of ϕ, and choosing a corresponding small value for δ. Both α and δ are input parameters to our algorithm. We use $\alpha = 0.2°$ and $\delta = 0.5°$. We choose a $\psi \in [\psi_A, \psi_B]$ when $|\psi_A - \psi_B| < \delta$. Further, our choice of ψ does not increase the RDC and RCSA RMSDs (i.e., the RMS deviation between the back-computed and experimental values) so much that they exceed user-defined thresholds; otherwise, the solution is discarded. Second, due to fine sampling of ϕ, often multiple pairs of (ϕ, ψ) cluster in a small region of the Ramachandran map. We cluster these solutions, and choose a set of representative candidates so that the complexity of the conformation tree search is not adversely affected.

A description of the core modules of RDC-CSA-ANALYTIC, and the inner working details are provided in the SI **Appendix B** available online [50].

The Analytic Step: Peptide Plane Orientations from N-H$^{\mathbf{N}}$ RDC, and ^{13}C-RCSA or ^{15}N-RCSA Measured in One Alignment Medium. To compute ψ_i for residue i, any of the ^{13}C′-RCSA, ^{15}N-RCSA or ^{1}H-RCSA can be used (see Table 1). Here we will use ^{13}C′-RCSA and derive the necessary mathematical tools for computing the dihedral ψ_i. Our derivation holds for ^{15}N-RCSA and ^{1}H-RCSA with minor modifications.

Proposition 1. *Given the diagonalized alignment tensor components S_{xx} and S_{yy}, the peptide plane P_i, the dihedral ϕ_i, and the ^{13}C′-RCSA $\Delta\delta_{C'}$ for residue i, there exist at most 4 possible values of the dihedral angle ψ_i that satisfy $\Delta\delta_{C'}$, and they can be computed exactly and in closed form by solving a quartic equation.*

Proof. The derivation below assumes standard protein geometry, which is exploited in the kinematics [52, 56]. Let the unit vector $\mathbf{v}_0 = (0, 0, 1)^T$ be the N-H$^{\mathrm{N}}$ bond vector of residue i in the local coordinate frame defined on the peptide plane P_i. Let $\mathbf{v}_1 = (x_1, y_1, z_1)^T$ and $\mathbf{v}_2 = (x_2, y_2, z_2)^T$ be the unit vectors defined with respect to the POF on the peptide plane P_{i+1}. We can write the forward kinematics relations between $\mathbf{v}_0$ and $\mathbf{v}_1$, and between $\mathbf{v}_0$ and $\mathbf{v}_2$ as follows:

$$\mathbf{v}_1 = \mathbf{R}_{i,\mathtt{POF}}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_m\, \mathbf{R}_z(\psi_i)\, \mathbf{R}_r\, \mathbf{v}_0, \tag{10}$$

$$\mathbf{v}_2 = \mathbf{R}_{i,\mathtt{POF}}\, \mathbf{R}_l\, \mathbf{R}_z(\phi_i)\, \mathbf{R}_m\, \mathbf{R}_z(\psi_i)\, \mathbf{R}'_r\, \mathbf{v}_0. \tag{11}$$

Here $\mathbf{R}_l, \mathbf{R}_m, \mathbf{R}_r$ and $\mathbf{R}'_r$ are constant rotation matrices. $\mathbf{R}_{i,\mathtt{POF}}$ is rotation matrix of P_i with respect to the POF. $\mathbf{R}_z(\phi_i)$ is the rotation about the z-axis by ϕ_i, and is a constant rotation matrix since ϕ_i is known. $\mathbf{R}_z(\psi_i)$ is the rotation about the z-axis by ψ_i. Let $c = \cos\psi_i$ and $s = \sin\psi_i$. Using this in Eq. (10) and Eq. (11) and simplifying we obtain

$$x_1 = A_{10} + A_{11}c + A_{12}s, \qquad x_2 = A_{20} + A_{21}c + A_{22}s, \tag{12}$$

$$y_1 = B_{10} + B_{11}c + B_{12}s, \qquad y_2 = B_{20} + B_{21}c + B_{22}s, \tag{13}$$

$$z_1 = C_{10} + C_{11}c + C_{12}s, \qquad z_2 = C_{20} + C_{21}c + C_{22}s, \tag{14}$$

where A_{ij}, B_{ij}, C_{ij} for $1 \leq i \leq 2$ and $0 \leq j \leq 2$ are constants. Using Eq. (12) to Eq. (14) in the reduced RCSA equation (Eq. (9)), and simplifying we obtain

$$K_0 + K_1c + K_2s + K_3cs + K_4c^2 + K_5s^2 = 0, \tag{15}$$

where K_i, $0 \leq i \leq 5$ are constants. Using half-angle substitutions

$$u = \tan(\frac{\psi_i}{2}),\ c = \frac{1-u^2}{1+u^2},\ \text{and}\ s = \frac{2u}{1+u^2} \tag{16}$$

in Eq. (15) we obtain

$$L_0 + L_1u + L_2u^2 + L_3u^3 + L_4u^4 = 0, \tag{17}$$

where L_i, $0 \leq i \leq 4$ are constants. Eq. (17) is a quartic equation that can be solved exactly and in closed form. For each real solution (at most four are possible), the corresponding ψ_i value can be computed using Eq. (16). □

Corollary 1. *Given the diagonalized alignment tensor components S_{xx} and S_{yy}, the peptide plane P_i, the dihedral ϕ_i, and a ψ-defining RDC r for P_{i+1}, there exist at most 4 possible values of the dihedral ψ_i that satisfy r. The possible values of ψ_i can be computed exactly and in closed form by solving a quartic equation.*

Proof. The proof follows from Proposition 1 by setting $\lambda_1 = 1$, $\lambda_2 = 0$ in Eq. (9), and treating $\mathbf{v}_1$ as the vector for which the ψ-defining RDC r is measured. □

2.4 The 3PLANAR Algorithm

We show that given any combination of three RCSAs and/or RDCs for internuclear vectors on a peptide plane (and in general, for any planar structural motif), there exist at most 16 possible orientations of the peptide plane that satisfy the three given orientational restraints. We further show that the 16 possible orientations can be computed in closed form by solving a *quadratic equation*. It is the only case where we have discovered a quadratic equation-based solution to constraints involving second-rank tensors, e.g., RDCs and RCSAs; all previous exact solutions to RCSA and/or RDC equations required solving quartic or higher degree equations. This we obtained by exploiting the symmetry of the equations in a novel way. Our main result is stated as the following proposition.

Proposition 2. *Given a rhombic alignment tensor, and 3 measurements, each of which is either an RCSA on a nucleus on the peptide plane P or an RDC on an internuclear vector on P, there exist at most 16 possible orientations for P that satisfy the 3 measurements, and these orientations can be written and solved in closed form by solving a quadratic equation.*

Proof. The proof is presented in the SI **Appendix C** available online [50]. □

Proposition 2, is incorporated into the 3PLANAR algorithm, which requires the following as input: (1) the diagonalized alignment tensor components (S_{yy}, S_{zz}); and (2) three orientational restraints, such as the CSA tensor parameters along with the RCSA values ($\delta_{xx}, \delta_{yy}, \Omega, \Delta\delta$) and/or RDCs. It outputs all the possible oriented peptide planes consistent with the RDC and/or RCSA data.

Table 2: Results on the alignment tensor computation, and RDC and RCSA data fit. (a) Experimental NMR data is from [12]. RMSD is the root-mean-square deviation between the back-computed and experimental values. (b) N-H^N RDC and $^{13}C'$-RCSA, and (c) N-H^N RDC and ^{15}N-RCSA, were used to compute the global fold. (d) The alignment tensors for the global folds computed by RDC-CSA-ANALYTIC agree well with that of the reference NMR structure.

Model	RDC and RCSA[a] used & RMSDs	Diagonalized Alignment Tensor S_{yy}, S_{zz}	Rhombicity[d] (ρ)
1D3Z	N-H^N: 1.11 Hz	-2.31, 51.17	0.61
1D3Z	N-H^N: 1.17 Hz, $^{13}C'$-RCSA: 6.85 ppb	-1.40, 50.57	0.63
1D3Z	N-H^N: 1.40 Hz, ^{15}N-RCSA: 10.08 ppb	-3.56, 49.40	0.57
RDC-CSA-ANALYTIC [b]	N-H^N: 1.21 Hz, $^{13}C'$-RCSA: 7.38 ppb	-0.71, 51.11	0.65
RDC-CSA-ANALYTIC [c]	N-H^N: 1.13 Hz, ^{15}N-RCSA: 9.22 ppb	-3.98, 46.10	0.55

3 Results and Discussion

3.1 Backbone Global Fold of Ubiquitin from Experimental RDC and RCSA

We applied our algorithm to compute the global fold of human ubiquitin. The protein ubiquitin has been a model system in many solution-state [6,7,12,14,24, 30,56] and solid-state [41,43] NMR studies. The solution structure of ubiquitin (PDB id: 1D3Z), and a 1.8 Å X-ray crystallographic structure of ubiquitin [55] (PDB id: 1UBQ), available in the PDB [2], were used as references. The experimental N-H^N RDC, $^{13}C'$-RCSA and ^{15}N-RCSA data were obtained from the previously published work by Cornilescu and Bax [12]. We used the uniform average values of the principal components of the CSA tensors reported in [12]. Such an assumption has been used widely in the literature for protein structure refinement against RCSA data [10, 26, 27, 53]. Whenever residue-specific CSA tensors can be determined, as in [6, 7, 30, 62, 63], RDC-CSA-ANALYTIC can use those tensors. The NOE restraints and hydrogen bond information for ubiquitin were extracted from the NMR restraint file for the PDB id 1D3Z [14].

Since RCSA measurement is usually accompanied by that of N-H^N RDC, recorded for the same sample under the same alignment conditions [10, 12, 28], computing an accurate backbone global fold from this limited amount of data, as a first step in protein structure computation, is of considerable interest. Here we present results of backbone global fold computation by RDC-CSA-ANALYTIC using (1) N-H^N RDC and $^{13}C'$-RCSA, and (2) N-H^N RDC and ^{15}N-RCSA.

The alignment tensor was computed from N-H^N RDC and $^{13}C'$-RCSA/^{15}N-RCSA data by bootstrapping the computation with an ideal helix for the helical region I23–K33 of ubiquitin (see the SI **Appendix B** [50]), and was used subsequently in the global fold computation. As summarized in Table 2, the alignment tensors agree well with those computed for the reference NMR structure. It is worth noting that if the alignment tensor is estimated by other methods [11], RDC-CSA-ANALYTIC can use that as input to compute the backbone global fold.

RDC-CSA-ANALYTIC computed accurate backbone conformations from N-H^N RDC plus either $^{13}C'$-RCSA or ^{15}N-RCSA data. As shown in Table 3, the backbone RMSDs between the computed SSEs and the reference structures are within 0.61 Å, and for about half of the cases they are less than 0.3 Å. The SSE back-

Table 3: Backbone RMSDs (Å) of SSE fragments computed by RDC-CSA-ANALYTIC. (a) NMR data is from [12]. (b) 12 H-bond information, and (c) 5 C^{α}-C^{α} approximate distance restraints derived from NOEs [56] were used.

Data Used; Reference[a]	α-helix I23–K33	β_1 Q2–T7	β_2 T12–V17	β_3 Q41–F45	β_4 K48–L50	β_5 S65–V70	β-sheet[b] $\beta_{1,\dots,5}$	Global Fold[c]
N-H^N,$^{13}C'$-RCSA; NMR	0.27	0.24	0.35	0.16	0.19	0.20	0.71	1.04
N-H^N,$^{13}C'$-RCSA; X-ray	0.23	0.32	0.37	0.28	0.20	0.25	0.79	1.09
N-H^N,^{15}N-RCSA; NMR	0.26	0.51	0.42	0.28	0.31	0.30	0.93	1.31
N-H^N,^{15}N-RCSA; X-ray	0.25	0.61	0.43	0.36	0.32	0.35	0.99	1.38

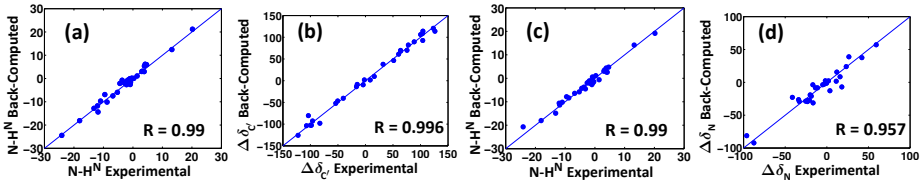

Fig. 2: Correlations between back-computed and experimental N-H^N RDCs and $^{13}C'$-RCSA (a, b), and those for N-H^N RDC and ^{15}N-RCSA (c, d) shown for the global folds computed by RDC-CSA-ANALYTIC.

bones computed using N-H^N RDC and $^{13}C'$-RCSA data agree better with the reference structures than those computed using N-H^N RDC and ^{15}N-RCSA data and compared with the reference structures. Table 2 and Figure 2 show that the back-computed RDCs and RCSAs for the RDC-CSA-ANALYTIC-computed structures are in good agreement with their experimental counterparts. For the structure computed using N-H^N RDC and ^{15}N-RCSA, the ^{15}N-RCSA Pearson's correlation coefficient is 0.957, and for other three cases the correlation coefficients are 0.99 or more (see Figure 2). This is explained by the slightly better quality structure obtained using N-H^N RDC and $^{13}C'$-RCSA data than that obtained using N-H^N RDC and ^{15}N-RCSA data. The β-sheet is computed using 12 hydrogen bond restraints in addition to the RDC and RCSA data. The α-helix (I23–K33) and the β-sheet for ubiquitin were packed using 5 approximate C^{α}-C^{α} distances derived from NOEs using the method described in [56]. The top 1000 packed structures obtained from the packing of ubiquitin α-helix and β-sheet, computed using N-H^N RDC and $^{13}C'$-RCSA data, have backbone RMSDs within the range 1.04–1.39 Å versus the reference NMR structure, and 1.09–1.42 Å versus the X-ray reference structure. The top 1000 packed structures obtained from the packing of ubiquitin α-helix and β-sheet, computed using N-H^N RDC and ^{15}N-RCSA data have backbone RMSDs within the range 1.31–1.86 Å versus the reference NMR structure, and 1.38–1.97 Å versus the X-ray reference structure. Figure 3 shows the overlay of the backbone fold computed by RDC-CSA-ANALYTIC versus the NMR and X-ray reference structures.

These results indicate that RDC-CSA-ANALYTIC can be used to compute accurate global folds from a minimal amount of RDC and RCSA data. Protein backbone global folds of similar resolution have been used successfully in empirical high-resolution structure determinations [65], NOE assignment [22,66], and side-chain resonance assignment [67]. Therefore, our method will be useful in high-

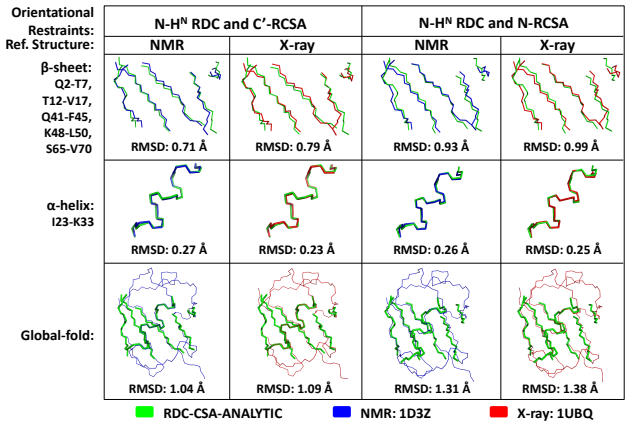

Fig. 3: Overlay of the ubiquitin global fold computed by RDC-CSA-ANALYTIC using N-H^N RDC and $^{13}C'$-RCSA or ^{15}N-RCSA versus the NMR and X-ray reference structures.

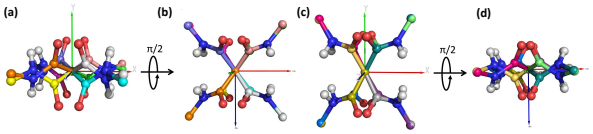

Fig. 4: The peptide plane orientations correspond to the two roots of the quadratic equations (Proposition 2) derived from C^α-C', C'-N and N-H^N RDCs measured in single alignment medium.

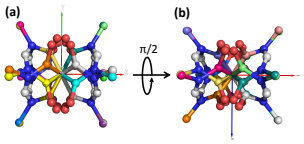

Fig. 5: Visualization of all sixteen peptide plane orientations together.

resolution protein structure determination. Furthermore, the use of RCSAs in the first stage of structure computation to compute accurate global folds is a novel concept, and our paper, being the first one to demonstrate this, can be a stepping stone to further research that exploits this new type of experimental data.

3.2 16-Fold Degeneracy of Peptide Plane Orientations

Our algorithm 3PLANAR was tested on the experimental RDC data for the protein ubiquitin (PDB id: 1D3Z) obtained from the BioMagResBank (BMRB) [54]. Using the singular value decomposition [29,56] module of RDC-CSA-ANALYTIC [52, 64, 65], we computed the principal components of the alignment tensor for ubiquitin using its NMR structure. We used C^α-C', C'-N and N-H^N RDCs, measured in one alignment medium, for the peptide plane defined by the residues Ala28 and Lys29 of ubiquitin. 3PLANAR then computed the 16 oriented peptide planes (individual planes are shown in the online SI **Appendix D** [50]). In Figure 4 (a, b) and (c, d), the two sets of 8 oriented peptide planes corresponding to the two roots of the quadratic equation are shown. Figure 5 shows the 16 oriented peptide planes visualized together. A counterclockwise rotation of 90° about the x-axis elucidates the symmetry in the peptide plane orientations. Similar results were obtained when N-H^N RDC, ^{15}N-RCSA and $^{13}C'$-RCSA data [12] was used, and the corresponding alignment tensor was computed by RDC-CSA-ANALYTIC.

3PLANAR is a $\Theta(1)$-time deterministic algorithm. During protein backbone structure determination, such as when using the RDC-ANALYTIC framework, the

multiple possible peptide plane orientations consistent with RDC/RCSA are usually ruled out by the kinematic coupling between peptide planes along the polypeptide chain, standard biophysical and protein geometry assumptions, or using additional experimental restraints.

4 Conclusions

We described a novel algorithm, RDC-CSA-ANALYTIC, that uses a sparse set of RCSAs and RDCs to compute the protein backbone global fold accurately. Our algorithm is the first algorithm to demonstrate that the orientational restraints from RCSA can be used in the initial stage of structure computation. We hope that this breakthrough will shed new light on the information content of RCSA, and help NMR structural biologists use our new ways of using RCSA to solve protein structures. Our algorithm barely scratches the surface of this new area, and much work remains to be done. Computing loop conformations using RCSA is an immediate future extension. Ubiquitin is the only protein for which we were able to obtain experimental ^{15}N-RCSA and ^{13}C$'$-RCSA data, available in the public domain. In future, we would like to test our algorithms on other protein systems, when experimental data becomes available for those systems.

When using orientational restraints in structure determination, it is important to know all the possible degeneracies associated with them, and their implications for structure determination. We gave exact and tight bounds on the orientational degeneracy of peptide planes computed using RDCs and/or RCSAs, and described a $\Theta(1)$-time algorithm, 3PLANAR, to compute them.

Although RDCs have been regularly used in protein structure determination, RCSAs have been used only in a few cases for structure validation [14] and refinement [10, 26, 27, 53]. We envision that algorithms, such as RDC-CSA-ANALYTIC, that use RCSA data plus RDCs during the initial stages of protein structure determination will play a large role in future.

References

1. M. Berjanskii and D. S. Wishart. *Nat Protoc*, 1:683–688, 2006.
2. H. M. Berman *et al.* *Nucleic Acids Res*, 28(1):235–242, 2000.
3. D. D. Boehr *et al.* *Science*, 313(5793):1638–1642, 2006.
4. G. Bouvignies *et al.* *Nature*, 477(7362):111–114, 2011.
5. A. T. Brünger. Yale University Press, New Haven, CT, 1992.
6. R. A. Burton and N. Tjandra. *J Biomol NMR*, 35(4):249–259, 2006.
7. R. A. Burton and N. Tjandra. *J Am Chem Soc*, 129:1321–1326, 2007.
8. J. Casey. *Proceedings of the Royal Society of London*, XIX:495–497, June 15, 1871.
9. A. Cavalli *et al.* *Proc Natl Acad Sci USA*, 104(23):9615–9620, 2007.
10. W.-Y. Choy *et al.* *J Biomol NMR*, 21:31–40, 2001.
11. G. Clore *et al.* *J Magn Reson*, 133(1):216 – 221, 1998.
12. G. Cornilescu and A. Bax. *J Am Chem Soc*, 122:10143–10154, 2000.
13. G. Cornilescu *et al.* *J Biomol NMR*, 13:289–302, 1999.
14. G. Cornilescu *et al.* *J Am Chem Soc*, 120:6836–6837, 1998.
15. B. B. Das *et al.* *J Am Chem Soc*, 134(4):2047–2056, 2012.

16. B. R. Donald. *Algorithms in Structural Molecular Biology*. The MIT Press, 2011.
17. B. R. Donald and J. Martin. *Prog NMR Spectrosc*, 55(2):101–127, 2009.
18. H. J. Dyson and P. E. Wright. *Nat Rev Mol Cell Biol*, 6:197–208, 2005.
19. A. W. Giesen *et al.* *J Biomol NMR*, 25:63–71, 2003.
20. M. Goldflam *et al.* *Methods in Molecular Biology*, 831:233–259, 2012.
21. A. Grishaev *et al.* *J Am Chem Soc*, 131(27):9490–9491, 2009.
22. P. Güntert. *Prog Nucl Magn Reson Spectrosc*, 43:105–125, 2003.
23. J.-C. Hus *et al.* *J Am Chem Soc*, 130:15927–15937, 2008.
24. O. F. Lange *et al.* *Science*, 320:1471–1475, 2008.
25. A. Leaver-Fay *et al.* *Methods Enzymol*, 487:545–574, 2011.
26. R. S. Lipsitz and N. Tjandra. *J Am Chem Soc*, 123(44):11065–11066, 2001.
27. R. S. Lipsitz and N. Tjandra. *J Magn Reson*, 164(1):171–176, 2003.
28. Y. Liu and J. H. Prestegard. *J Biomol NMR*, 47(4):249–258, 2010.
29. J. A. Losonczi *et al.* *J Magn Reson*, 138:334–342, 1999.
30. K. Loth *et al.* *J Am Chem Soc*, 127(16):60626068, 2005.
31. S. C. Lovell *et al.* *Proteins: Struct, Funct, Genet*, 40:389–408, 2000.
32. T. G. Oas *et al.* *J Am Chem Soc*, 109(20):5962–5966, 1987.
33. K. Pervushin *et al.* *Proc Natl Acad Sci USA*, 94(23):12366–12371, 1997.
34. J. H. Prestegard *et al.* *Chemical Reviews*, 104:3519–3540, 2004.
35. S. Raman *et al.* *Science*, 327:1014–1018, 2010.
36. T. L. Religa *et al.* *Nature*, 437(7061):1053–1056, Oct 13 2005.
37. G. Salmon. Longmans, Green and Company, London, 1912.
38. L. Salmon *et al.* *J Am Chem Soc*, 132(24):8407–8418, 2010. PMID: 20499903.
39. C. R. Sanders II and G. C. Landis. *J Am Chem Soc*, 116(14):6470–6471, 1994.
40. A. Saupe. *Angewandte Chemie*, 7(2):97–112, 1968.
41. P. Schanda *et al.* *J Am Chem Soc*, 132(45):15957–15967, 2010.
42. C. D. Schwieters *et al.* *J Magn Reson*, 160:65–73, 2003.
43. K. Seidel *et al.* *ChemBioChem*, 6(9):1638–1647, 2005.
44. Y. Shen *et al.* *J Biomol NMR*, 44:213–223, 2009.
45. Y. Shen *et al.* *Proc Natl Acad Sci USA*, 105(12):4685–4690, 2008.
46. D. Sitkoff and D. A. Case. *Prog Nucl Magn Reson Spectrosc*, 32(2):165–190, 1998.
47. F. Tian *et al.* *J Am Chem Soc*, 123:11791–11796, 2001.
48. N. Tjandra and A. Bax. *Science*, 278:1111–1114, 1997.
49. J. R. Tolman *et al.* *Proc Natl Acad Sci USA*, 92:9279–9283, 1995.
50. C. Tripathy, A. K. Yan, P. Zhou, and B. R. Donald. Supporting information: `http://www.cs.duke.edu/donaldlab/Supplementary/recomb13/` . 2013.
51. C. Tripathy *et al.* In *Proceedings of RECOMB*, LNBI/LNCS 6577:483–498, 2011.
52. C. Tripathy *et al.* *Proteins: Struct, Funct, Bioinf*, 80(2):433–453, 2012.
53. V. Tugarinov *et al.* *Proc Natl Acad Sci USA*, 102(3):622–627, 2005.
54. E. L. Ulrich *et al.* *Nucleic Acids Res*, 36(Database issue):D402–D408, 2008.
55. S. Vijay-kumar *et al.* *J Mol Biol*, 194:531–44, 1987.
56. L. Wang and B. R. Donald. *J Biomol NMR*, 29(3):223–242, 2004.
57. L. Wang and B. R. Donald. In *Proceedings of CSB*, pages 189–202, 2005.
58. L. Wang *et al.* *J Comput Biol*, 13(7):1276–1288, 2006.
59. D. S. Wishart and D. A. Case. *Methods Enzymol*, 338:3–34, 2002.
60. B. J. Wylie *et al.* *J Am Chem Soc*, 127(34):11946–11947, 2005.
61. B. J. Wylie *et al.* *J Am Chem Soc*, 131(3):985–992, 2009.
62. L. Yao *et al.* *J Am Chem Soc*, 132(31):10866–10875, 2010.
63. L. Yao *et al.* *J Am Chem Soc*, 132(12):4295–4309, 2010.
64. A. Yershova *et al.* In *Proceedings of WAFR*, 68:355–372, 2010.
65. J. Zeng *et al.* *J Biomol NMR*, 45(3):265–281, 2009.
66. J. Zeng *et al.* In *Proceedings of CSB*, pages 169–181. ISBN 1752–7791, 2008.
67. J. Zeng *et al.* *J Biomol NMR*, 50(4):371–395, 2011.
68. J. Zeng *et al.* *J Biomol NMR*, pages 1–14, 2012.